Letter to the Editor

Am. J. Hum. Genet. 67:1355, 2000

Reply to Kong and Nicolae

To the Editor:

We thank Kong and Nicolae (2000) for their insightful discussion of our proposed randomization procedure for linkage analysis (Zhao et al. 1999). In light of the example in their discussion, we agree that our proposed method is anticonservative for nuclear families when both parents are missing. For families of other structures, Kong and Nicolae stated that "it can be shown that, at least for the single marker case, it is asymptotically slightly conservative for sib-pair data with genotypes on both parents." They further stated that, "In general, with complete descent information, the randomization procedure gives valid exact *P* values that are the same as those obtained by direct simulation and the 'exact *P* values' of GENEHUNTER (Kruglyak et al. 1996)." We agree that, with complete descent information, the randomization procedure gives valid statistical inference. However, we do not think that the results from the randomization procedure are the same as those obtained by direct simulation and the "exact *P* values" of GENEHUNTER. In this letter, we use examples to illustrate the differences between our proposed randomization procedure and the two alternative methods—direct simulation and the perfect data approximation in GENEHUNTER—in the determination of statistical significance for genetic linkage.

For any direct simulation method, a crossover process model must be specified to describe the distribution of the recombination events along the chromosomes during meiosis. Because crossover interference has been shown to exist in humans (e.g., see Broman and Weber 2000), direct simulations should be based on a model that can incorporate crossover interference—for example, the χ^2 model (Zhao et al. 1995)—instead of the more commonly used Poisson model, which assumes the absence of crossover interference. However, the appropriateness of such crossover process models needs to be tested using extensive empirical data. Moreover, the effect of model misspecifications cannot be determined. On the other hand, the randomization procedure proposed in our article depends only on the observed recombination patterns rather than on a particular crossover model. Consider a family with one child, his or her two parents, and all four grandparents. For each marker, the inheritance vector for the child has two components (*f*,*m*), where $f = 0$ or 1 if the grandpaternal or grandmaternal allele was transmitted to the child from his/her father and $m = 0$ or 1 if the grandpaternal or grandmaternal allele was transmitted to the child from his/her mother. Assume the most ideal case, in which we can identify the grandparental origin for the two chromosomes in the child for all genetic markers being studied—that is, we can uniquely determine the inheritance vector of the child for all markers; therefore, (*f*,*m*) is known without ambiguity. In this case, we can pull the *f* component in the inheritance vectors for all the markers into a vector to summarize the transmissions from the father to the child across all the markers and the *m* component for all the markers in a separate vector to represent the transmission from the mother to the child across all the markers. For example, consider 10 markers and the following two vectors representing transmissions from the father and the mother, respectively, to the child: (1,1,0,0,0,0,0,1,1,1) and (0,0,0,0,1,1,1,1,1,1). For this example, the child inherited the grandmaternal alleles from the father at markers 1, 2, 8, 9, and 10 and the grandpaternal allele from the father at markers 3–7. Similarly, the child inherited the grandmaternal alleles from the mother at markers 5–10 and the grandpaternal allele from the mother at markers 1–4. Under the randomization procedure proposed in our article, for the 10 markers for this child, it is equally likely that each randomization would generate the following four inheritance vector pairs: (a) (1,1,0,0,0,0,0,1,1,1) and (0,0,0,0,1,1,1,1,1,1); (b) (0,0,1,1,1,1,1,0,0,0) and (0,0,0,0,1,1,1,1,1,1); (c) (1,1,0,0,0,0,0,1,1,1) and (1,1,1,1,0,0,0,0,0,0); and (d) (0,0,1,1,1,1,1,0,0,0) and (1,1,1,1,0,0,0,0,0,0). Therefore, the number of recombination events and the distribution of the recombinations are preserved in each randomized sample, and no specific crossover process models are used in the simulations. In contrast, for direct simulation methods, the number and positions of recombination events will differ across simulations.

Consider a family with two parents and two affected children. Using the notation by Kong and Nicolae (2000), we distinguish four states, for this pedigree,

among the two affected children: (0,0) corresponds to the two sibs sharing zero alleles identical by descent (IBD); (1,1) corresponds to sharing both alleles IBD; (1,0) corresponds to IBD sharing of the paternal allele but not the maternal allele; and (0,1) corresponds to not sharing the paternal allele but sharing the maternal allele. Assume that all four individuals in the pedigree have been genotyped at a single genetic marker and that the father has genotype (*A,A*), the mother has genotype (*B,C*), the first affected child has genotype (*A,B*), and the second affected child has genotype (*A,B*). Because the father is homozygous at this marker, we cannot uniquely determine the number of alleles IBD between the two affected children. With the notation defined by Kong and Nicolae (2000), for this pedigree $p(1,1) = p(0,1) = 1/2$ and the nonparametric linkage analysis score is $1/2 \times 1 + 1/2 \times 2 = 1.5$. The randomization procedure would generate the following four sets of probabilities with equal chance: (a) $\{p(0,0) = 0, p(0,1) = 1/2, p(1,0) = 0, p(1,1) = 1/2\}$; (b) $\{p(0,0) = 0, p(0,1) = 1/2, p(1,0) = 1/2, p(1,1) = 0\}$; (c) $\{p(0,0) = 1/2, p(0,1) = 0, p(1,0) = 1/2, p(1,1) = 0\}$; and (d) $\{p(0,0) = 1/2, p(0,1) = 0, p(1,0) = 1/2, p(1,1) = 0\}$. Therefore, in the randomized sample, the test statistic NPL = .5 and 1.5 with equal probability, whereas the "exact *P* value" in GENEHUNTER is calculated by means of a different reference distribution, in which the NPL = 0, 1, and 2 with probability 1/4, 1/2, and 1/4, respectively. Therefore, the procedure in GENEHUNTER-PLUS overestimates the variance for the NPL statistic for this particular family. In fact, this conservative approach of the statistical significance level evaluation in GENEHUNTER was the motivation of a likelihood-based approach in GENEHUNTER-PLUS by Kong and Cox (1997).

As a final note, the families analyzed in the insulin-dependent diabetes mellitus data set in our article have both parents available. Therefore, for this particular data set, the differences between the results from GENEHUNTER-PLUS and the randomization procedure are not likely to be due to the bias caused by incomplete parental information in the data set.

HONGYU ZHAO,[1,2] KATHLEEN R. MERIKANGAS,[1] AND KENNETH K. KIDD[2]
Departments of [1]Epidemiology and Public Health and [2]Genetics
Yale University School of Medicine
New Haven

References

Broman KW, Weber JL (2000) Characterization of human crossover interference. Am J Hum Genet 66:1911–1926

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

Kong A, Nicolae DL (2000) On a randomization procedure. Am J Hum Genet 67:1352–1355 (in this issue)

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Zhao H, Merikangas KR, Kidd KK (1999) On a randomization procedure in linkage analysis. Am J Hum Genet 65: 1449–1456

Zhao H, Speed TP, McPeek MS (1995) Statistical analysis of crossover interference using the chi-square model. Genetics 139:1045–1056

Address for correspondence and reprints: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

0002-9297/2000/6705-0036$02.00